# EXPLOITING COMBINATORIAL COMPLEXITY – SEARCHING FOR NEW FUNCTIONAL ENTITIES IN THE CELL

Jens Hollunder*, Andreas Beyer, Thomas Wilhelm
Theoretical Systems Biology
Institute of Molecular Biotechnology
D-07745 Jena, Beutenbergstr. 11.

*Abstract*

Complex cellular processes are accomplished by the concerted action of hierarchically organized functional modules. Proteins associate to build protein complexes which act as highly specialized molecular machines. We present a statistical procedure to find insightful sub-structures in protein complexes based on large-scale protein complex data: we identify statistically significant common protein subcomplexes (SCs) representing more reliable protein assemblies than the original complexes. Different protein complex data sets of two model organisms *Saccharomyces cerevisiae* and *Escherichia coli* are separately analyzed. On the one hand, we identify well-characterized protein machineries with known functions. On the other hand, we also identify hitherto unknown protein assemblies consisting of either unknown proteins or proteins with different functional annotations. Finally, we demonstrate unique properties of subcomplex proteins in these two model organisms that underline the distinct roles of SCs: (i) subcomplexes are enriched with essential proteins (which implies that SC proteins are evolutionary strongly conserved), (ii) expression of subcomplex proteins is higher correlated than that of other proteins, (iii) SCs are functionally more homogeneous than the experimentally found protein complexes. The latter property is exploited to propose functions for so far unknown proteins of *E. coli* and *S. cerevisiae*.

*Keywords*

Proteome, Protein complex, Protein-protein interaction, Gene annotation, *E. coli*, *S. cerevisiae*.

## Introduction

Complex cellular processes are accomplished by the concerted action of hierarchically organized functional modules. Multi-protein complexes comprise the most versatile modules and cellular complexity heavily relies on the contextual combination of gene products (Gavin and Superti-Furga, 2003; Wilhelm et al., 2003; de Lichtenberg et al., 2005). These assemblies represent more than the sum of their parts. Deciphering the whole genome is much easier than understanding the complete proteome/ translatome. During the last decades many protein complexes have been identified by detailed individual analyses. In recent years large-scale identification of protein complexes became feasible. The first systematic analyses of protein complexes have been published three years ago (Gavin et al., 2002 and Ho et al., 2002). By using similar pull-down procedures both groups identified hundreds of different protein complexes in the model organism *Saccharomyces cerevisiae*. The only other large-scale identification of protein complexes appeared just recently: Butland et al. (2005) studied protein complexes in *E. coli*.

Up to now, all computational approaches to detect protein complexes and functional modules are based on

* Jens Hollunder, Beutenbergstr. 11, D-07745 Jena, Germany, Fax: +49-3641 65 6191, E-mail: hollund@imb-jena.de

binary protein-protein interaction networks, but it is known that protein complex assembly involves cooperative binding and cannot easily be mapped by binary interactions (Gavin and Superti-Furga, 2003).

Here, we report a method to systematically identify all statistically significant protein subcomplexes in a protein complex data set. No additional information is needed for this approach. The corresponding protein-protein interactions are of higher credibility than other interactions in a protein complex. We demonstrate unique properties of subcomplexes and subcomplex proteins for *S. cerevisiae* (Hollunder et al., 2005) and *E. coli*.

## Methods

The algorithm consists of two steps: (i) counting the frequency of subcomplexes occurring in the measurements ('occurrence value') and (ii) identifying statistically significant subcomplexes (p-values). Note that we account only for complexes consisting of different proteins, i.e. homodimers are treated as one protein, and a complex consists of at least two different proteins. This is due to the fact that current high-throughput methods are unable to determine the stoichiometry of the complexes.

At first, we count the occurrence of all possible combinations of different proteins in each data set. By 'host complex' we denote measured complexes containing a certain subcomplex. The occurrence number of a subcomplex is the number of host complexes in the original data set (measured complexes) containing the subcomplex. Hence, a large occurrence number corresponds to a large number of measurements confirming the existence of this subcomplex.

Statistical significance of the subcomplexes is calculated by a comparison of the SC occurrence values with occurrence values of appropriately randomized data sets. At first we analytically calculate the average probability of a subcomplex in the respective random data set. Based on this probability we determine the p-value of the observed occurrence assuming a binomial distribution.

## Results

### *(I) Saccharomyces cerevisiae*

*Data used*

We analyze four different protein complex data sets of the yeast *S. cerevisiae*: HMS-PCI: obtained by high-throughput mass spectrometric protein complex identification (548 different protein complexes, Ho et al., 2002); TAP$_{\text{raw}}$: all TAP[1] purified complexes (446 different protein complexes); TAP$_{\text{cur}}$: manually curated TAP[1] protein complexes (229 different protein complexes); MIPS: the manually curated yeast protein complex data of the Munich information center for protein sequences (214 different protein complexes, Mewes et al., 2004, MIPS updated 2004).

*Subcomplexes*

We extracted 1.468 (HMS-PCI), 1.238 (TAP$_{\text{raw}}$), 519 (TAP$_{\text{cur}}$) and 32 (MIPS) SCs from the four data sets of *Saccharomyces cerevisiae* (see *Supporting Information*, http://www.imb-jena.de/tsb/s_cerevisiae). For the analysis we distinguish known and unknown SCs: Known subcomplexes are protein assemblies that have already been described in the literature with a similar protein composition as the subcomplex identified here, e.g. eIF2B, Casein kinase II, TAF scaffolding factor, 19/22S regulator of the proteasome, and RNA polymerase III (Hollunder et al., 2005).

Furthermore, we also identified several previously undescribed protein assemblies which could act as functionally important entities in the cell.

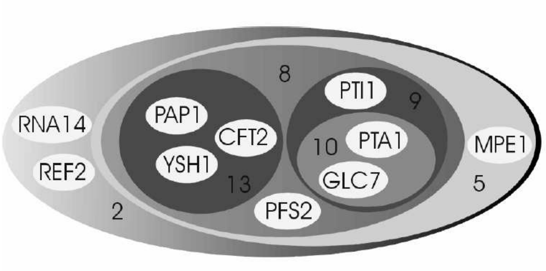

*Figure 1. Modular composition of protein sub-complexes (SCs); example from S. cerevisiae.*

Some of the SCs have a rather intricate modular composition consisting of separate smaller modules. Figure 1 shows an example of a hierarchically organized protein complex consisting of partly independent modules. The subcomplex Cft2/Ysh1/Pap1 is found 13 times in TAP$_{\text{raw}}$, which renders this interaction triple highly significant. It contains half of the pre-mRNA 3'-end processing factor CFII (Zhao et al.) and interestingly, a further protein Pap1 (poly(A) polymerase). The two SCs Pti1/Glc7/Pta1 and Cft2/Ysh1/Pap1 together with Pfs2 form a significant subcomplex of size 7. Pfs2 is known to serve as a bridging protein for mRNA processing factors (Ohnacker et al. 2000). The significance of these subcomplexes suggests that these combinations act as distinct functional entities in the cell. Subcomplexes with the additional mRNA processing proteins Mpe1, Ref2 and Rna14 are also statistically significant.

*Properties of protein subcomplexes*

SCs represent more reliable protein assemblies than the original measured complexes. They have special unique properties which underline the distinct role of SCs (Hollunder et al., 2005).

SCs are characterized by a more homogeneous spatial and functional composition than the host complexes. This

[1] TAP: tandem-affinity purification data (Gavin et al., 2002) give rise to the two different data sets TAP$_{\text{raw}}$ and TAP$_{\text{cur}}$

higher confidence allows us to propose functions and subcellular localizations for not yet annotated proteins.

The importance of protein assemblies is underlined by the essentiality of the contained proteins. In SCs the fraction of essential proteins is approximately 70% which shows that these assemblies play a central role. Moreover, SCs are enriched with synthetic lethal protein pairs. Additionally, we find that mutations in subcomplex proteins have higher fitness effects than mutations in other proteins. The abundance of SC proteins is less variable compared to other proteins.

### *(II) Escherichia coli*

#### *Data used*

We analyze the recently published sole protein complex data set of *E. coli* (Butland et al., 2005). It contains 530 different protein complexes with 1.289 different proteins. Starting from this data set we calculated another protein complex data set: for the "*E. coli-strong*" data set we only counted interactions where a bait protein "A" catches a protein "B" and protein "B" as a bait protein catches protein "A". This reduced yet highly reliable data set contains 145 different protein complexes with 145 different proteins.

Functional annotation was taken from Blattner et al. (1997) and essentially data from Gerdes et al. (2003), who identified 630 essential genes by using a genetic footprinting technique for a genome-wide assessment of genes. The co-expression information was obtained from Chang et al. (2002). They studied genes that are involved in growth phase transitions and divided the expression data into five clusters.

*Table 1. Properties of E. coli proteins in complexes and subcomplexes. Numbers in parenthesis correspond to the most significant subcomplexes (p-value ≤ 1E-200). The functional homogeneity shows the average homogeneity of all protein complexes and subcomplexes. Essentiality shows the fraction of essential genes. Co-expression shows the fraction of complexes and subcomplexes in which more than 50% of the proteins belong to the same expression cluster.*

| | functional homogeneity | essentiality | co-expression |
|---|---|---|---|
| complexes | 0.31 | 0.24 | 0.05 |
| subcomplexes | 0.93 (0.98) | 0.38 (0.47) | 0.34 (0.35) |

#### *Subcomplexes*

By using our method we obtained 80.481 different SCs for *E. coli* (see *Supporting Information,* http://www.imb-jena.de/tsb/e_coli). 10.527 SCs are statistically highly significant (p-value ≤ 1E-200). The "*E. coli-strong*" data set yielded 152 subcomplexes.

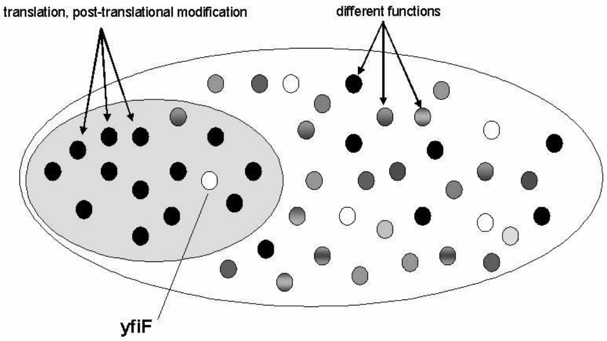

*Figure 2. Functional homogeneity of protein subcomplexes (SCs); example from E. coli. Each circle represents a protein, with the grey scales denoting different functional groups. Proteins with unknown functions are represented by an open circle. The highlighted region indicates a significant subcomplex (p-value ≤ 1E-200). The protein yfiF is not yet annotated and most likely shares the function of the SC.*

#### *Properties of protein subcomplexes*

We show that *E. coli* SCs have the same unique properties as *S. cerevisiae* SCs.

Table 1 summarizes a number of features of SCs in *E. coli*. The higher functional homogeneity of SCs allows us to propose functions for so far unknown proteins: unknown proteins belonging to highly significant subcomplexes (p-value ≤ 1E-200) with a homogenous functional composition (>80% of the proteins have one function) get assigned the function of this subcomplex.

Figure 2 shows an example of a measured protein complex. It demonstrate that the SC is functionally much more homogeneous than the host complex. This is a common property of most SCs.

We proposed new functions for 10 hitherto unknown proteins (see Supporting Information, http://www.imb-jena.de/tsb/e_coli).

Essentially of proteins can be used to measure the importance of protein assemblies. 16% of all *E. coli* proteins and about 24% of the proteins in protein complexes are essential. The fraction of essential proteins in SCs rises to approximately 40%, highlighting the importance of SCs in *E. coli*. The last column (co-expression) in Table 1 confirms that more than one quarter of the SCs are composed of proteins with more than 50% belonging to the same expression cluster (Chang et al., 2002).

## Conclusions

We have developed a method to identify protein assemblies as subcomplexes of larger protein complexes that may act as relatively independent functional entities in the cell.

In *S. cerevisiae* we identified already known subcomplexes which give confidence to the applicability of the method. Additionally, we also identified previously unknown protein assemblies that may fulfill special cellular functions (Hollunder et al., 2005).

We found unique properties for SCs and subcomplex proteins with respect to their essentiality, co-expression, and functional homogeneity. The higher confidence of protein interactions in SCs is exploited to propose functions for so far unknown proteins (20 in *S. cerevisiae*. and 10 in *E. coli*).

Recently, Butland et al. (2005) measured protein complexes in *E. coli* which seemingly have a higher reliability than measured protein complexes in *S. cerevisiae*. Interestingly, in this new *E. coli* protein complex data set we identified much more protein subcomplexes than in yeast.

More than 800 subcomplexes are exclusively composed of proteins belonging to one single expression cluster: All these SCs contain proteins that are downregulated during the five growth transitions, including genes involved in transcription, translation, biosynthesis and energy metabolism (Chang et al., 2002).

One third of the "*E. coli-strong*" complexes are also found as highly significant protein subcomplexes (p-value ≤ 1E-200) in the full data set.

The identification of functional protein assemblies as 'molecular machines' (Gavin and Superti-Furga, 2003) remains one of the most important tasks for future biology.

## Acknowledgements

This work has been funded by the Federal Ministry of Education and Research, Germany (Grand 0312704E).

## References

Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, *277*, 1473.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., Emili, A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, *433*, 531.

Chang, D.-E., Smalley D. J., Conway, T. (2002). Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model. *Mol. Microbiology*, *45*, 289.

de Lichtenberg, U., Jensen, L. J., Brunak, S., Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, *307*, 724.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*, 141.

Gavin, A.-C., Superti-Furga, G. (2003). Protein complexes and proteome organization from yeast to man. *Curr. Opin. Chem. Biol.*, *7*, 21.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A.-L., Oltavai, Z. N., Osterman A. L. (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J. Bacteriology*, *185*, 5673.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, *415*, 180.

Hollunder, J., Beyer, A., Wilhelm, T. (2005). Identification and characterization of protein subcomplexes in yeast. *Proteomics*, *5*, 2082.

Mewes H. W., Amid C., Arnold R., Frishman D., Guldener U., Mannhaupt G., Munsterkotter M., Pagel P., Strack N., Stumpflen V., Warfsmann J., Ruepp A., (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, *32*, D41.

Ohnacker, M., Barabino S. M., Preker P. J., Keller, W., (2000). The WD-repeat protein Pfs2p bridges two essential factors within the yeast pre-mRNA 3'-end-processing complex. *EMBO J.*, *19*, 37.

Wilhelm, T., Nasheuer, H.-P., Huang, S., (2003). Physical and functional modularity of the protein network in yeast. *Mol. Cell. Prot.*, *2*, 292.

Zhao, J., Kessler, M., Helmling, S., O´Connor J. P., Moore, C. (1999). Pta1, a component of yeast CFII, is required for both cleavage and poly(A) addition of mRNA precursor. *Mol. Cell. Biol.*, *19*, 7733.